Measuring Conditional Dependence with Kernels

Kenji Fukumizu
Institute of Statistical Mathematics

Joint work with A. Gretton, X. Sun, and B. Schölkopf

2nd Workshop on Machine Learning and Optimization
October 12, 2007 @ ISM

Outline

1. Introduction

2. Characterization of conditional independence with kernels

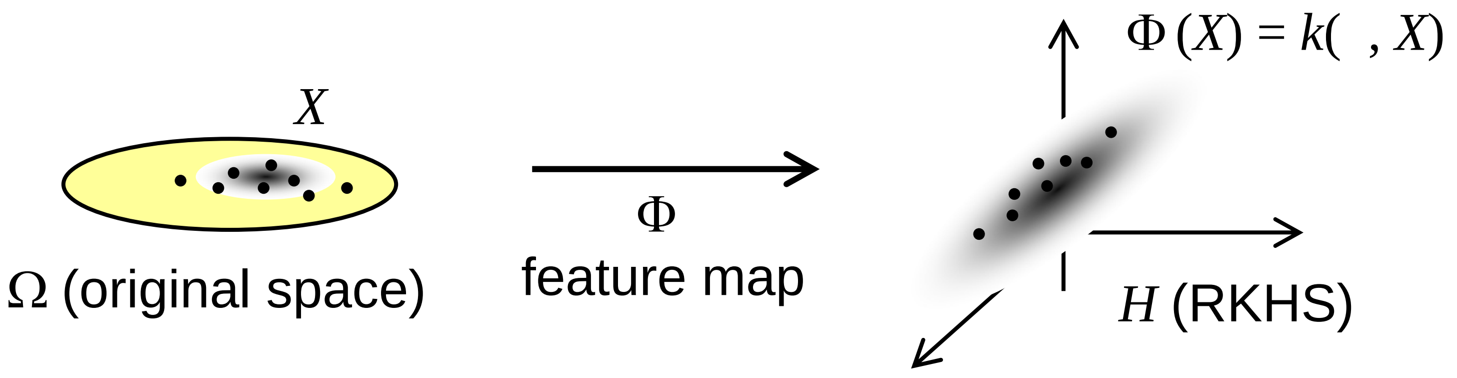
3. Conditional dependence measure with normalized operators and its kernel-free expression.

4. Experiments

5. Concluding remarks

Introduction

- "Kernel methods" for nonlinear relations
  - Positive definite kernels have been used for capturing nonlinearity of original data.   *e.g*. Support vector machine.
  - Kernelization: mapping data into a functional space (RKHS) and apply linear methods on RKHS.
  - Consider linear statistics (mean, variance, …) on RKHS, and their meaning on the original space.



- Representing probabilities
  - Determining probabilities (Arthur Gretton's talk)
  - Characterizing independence (Arthur Gretton's talk)
  - Characterizing conditional independence

- Motivation
  - Dependence among many variables
  - Conditional independence is essential for many probabilistic modeling
    e.g. graphical modeling

Positive Definite Kernel and RKHS

- Positive definite kernel (p.d. kernel)

Ω: set. $k:\Omega\times\Omega\to\mathbf{R}$

k is positive definite if $k(x,y) = k(y,x)$ and for any $n\in\mathbf{N}$, $x_1,\ldots x_n\in\Omega$ the matrix $\left(k(x_i,x_j)\right)_{i,j}$ (Gram matrix) is positive semidefinite.

  – Example: Gaussian RBF kernel $k(x,y)=\exp\left(-\|x-y\|^2/\sigma^2\right)$

- Reproducing kernel Hilbert space (RKHS)

k: p.d. kernel on Ω.

$\Rightarrow$ $\exists 1\ H$: reproducing kernel Hilbert space (RKHS)

1) $k(\cdot,x)\in H$ for all $x\in\Omega$.

2) $\mathrm{Span}\{k(\cdot,x)\mid x\in\Omega\}$ is dense in H.

3) $\langle k(\cdot,x), f\rangle_H = f(x)$ (reproducing property)

Functional data (feature map)

$$\Phi : \Omega \to H, \quad x \mapsto k(\cdot, x) \qquad i.e. \quad \Phi(x) = k(\cdot, x)$$

$$\langle \Phi(x), f \rangle = f(x)$$ (reproducing property)

Data: $X_1, \ldots, X_N$ → $\Phi_X(X_1), \ldots, \Phi_X(X_N)$: functional data

Why RKHS?

– By the reproducing property, computation of the inner product on RKHS does not need expansion by basis functions.

$$f(\cdot) = \sum_i a_i k(\cdot, x_i), \quad g(\cdot) = \sum_j b_j k(\cdot, x_j)$$

$$\Rightarrow \quad \langle f, g \rangle = \sum_{i,j} a_i b_j k(x_i, x_j)$$

Advantageous for high-dimensional data of small sample size.

Representing Nonlinear Dependence

- **Kernel Statistics: linear statistics on RKHS**

  X , Y : general random variables on Ω_X and Ω_Y , resp.

  Prepare RKHS (H_X, k_X) and (H_X , k_X) defined on Ω_X and Ω_Y, resp

  Define random variables on the RKHS H_X and H_Y by

$$\Phi_X(X) = k_X(\cdot, X) \qquad \Phi_Y(Y) = k_Y(\cdot, Y)$$

  – Covariance

$$\Sigma_{YX} \equiv E[(\Phi_Y(Y) - \mu_X)(\Phi_X(X) - \mu_Y)^T] \longrightarrow \Sigma_{XY} = O \Leftrightarrow X \perp\!\!\!\perp Y$$

  – Conditional covariance

$$\Sigma_{YX|Z} \equiv \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX} \longrightarrow \Sigma_{YX|Z} = O \Leftrightarrow X \perp\!\!\!\perp Y \mid Z$$

  – c.f. Gaussian variables

$$V_{XY} = O \Leftrightarrow X \perp\!\!\!\perp Y$$

$$V_{YX|Z} = O \Leftrightarrow X \perp\!\!\!\perp Y \mid Z$$

Richness Assumption on RKHS

k: kernel on a measurable space $(\Omega, \mathcal{B})$. H: associated RKHS.

Assumption (A):

$\exists q \geq 1.$ $H + \mathbf{R}$ is dense in $L^q(P)$ for any probability P on $(\Omega, \mathcal{B})$,

– RKHS can approximates various functions such as the index function of a measurable set, polynomials, and $e^{\sqrt{-1}\omega^T x}$.

– Example: Gaussian kernel on the entire $\mathbf{R}^m$

$$k_G(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2} \right)$$

Laplacian kernel on the entire $\mathbf{R}^m$

$$k_L(x, y) = \exp\left(-\lambda \sum_{i=1}^m |x_i - y_i| \right)$$

Covariance on RKHS

– Definition: cross-covariance operator

X , Y : general random variables on Ω_X and Ω_Y , resp.

Prepare RKHS (H_X, k_X) and (H_X , k_X) defined on Ω_X and Ω_Y, resp.

There is a unique operator $\Sigma_{YX} : H_X \to H_Y$ such that

$$\langle g, \Sigma_{YX} f \rangle = E[g(Y) f(X)] - E[g(Y)]E[f(X)] \quad (= \mathrm{Cov}[f(X), g(Y)])$$

for all $f \in H_X, g \in H_Y$

– Independence by cross-covariance operator

Under (A),

X and Y are independent $\Leftrightarrow$ $\Sigma_{XY} = O$

$$E[g(Y) f(X)] = E[g(Y)]E[f(X)]$$

- *c.f.* Characteristic function

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad E_{XY}[e^{\sqrt{-1}(uX+vY)}] = E_X[e^{\sqrt{-1}uX}]E_Y[e^{\sqrt{-1}vY}]$$

Conditional Covariance on RKHS

- Conditional Cross-covariance operator

  X, Y, Z : random variables on $\Omega_X, \Omega_Y, \Omega_Z$ (resp.).

  $(H_X, k_X), (H_Y, k_Y), (H_Z, k_Z)$: RKHS defined on $\Omega_X, \Omega_Y, \Omega_Z$ (resp.).

  – Conditional cross-covariance operator $H_X \to H_Y$

$$\Sigma_{YX|Z} \equiv \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$$

  – Conditional covariance operator

$$\Sigma_{YY|Z} \equiv \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$$

  – Note: Σ_{ZZ}^{-1} may not exist. But, we have the decomposition

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} W_{YX} \Sigma_{XX}^{1/2} \qquad \text{with operator norm } \|W_{YX}\| \le 1$$

  Rigorously, define $\Sigma_{YX|Z} \equiv \Sigma_{YX} - \Sigma_{YY}^{1/2} W_{YZ} W_{ZX} \Sigma_{XX}^{1/2}$

Relation with regression error

Theorem (FBJ'06)

Y, Z : random variables on Ω_Y, Ω_Z (resp.).

(H_Y, k_Y), (H_Z, k_Z) : RKHS defined on Ω_Y, Ω_Z (resp.).

$$\langle g, \Sigma_{YY|Z} g \rangle = \inf_{f \in H_Z} E\left| \left(g(Y) - E[g(Y)] \right) - \left(f(Z) - E[f(Z)] \right) \right|^2$$

$$= \inf_{f \in H_Z} Var\left[g(Y) - f(Z) \right] \qquad (\forall g \in H_Y)$$

c.f. for Gaussian variables,

$$b^T V_{YY|Z} b = \min_a \left| b^T \tilde{Y} - a^T \tilde{Z} \right|^2 \qquad (\tilde{Y} = Y - E[Y], \tilde{Z} = Z - E[Z])$$

Residual error of linear regression is given by the conditional covariance matrix.

– Rough sketch of the proof

$$E\left|\left(g(Y)-E[g(Y)]\right)-\left(f(Z)-E[f(Z)]\right)\right|^2$$

$$=\langle f,\Sigma_{ZZ}f\rangle-2\langle f,\Sigma_{ZY}g\rangle+\langle g,\Sigma_{YY}g\rangle$$

$$=\left\|\Sigma_{ZZ}^{1/2}f\right\|^2-2\left\langle f,\Sigma_{ZZ}^{1/2}W_{ZY}\Sigma_{YY}^{1/2}g\right\rangle+\left\|\Sigma_{YY}^{1/2}g\right\|^2$$

$$=\left\|\Sigma_{ZZ}^{1/2}f-W_{ZY}\Sigma_{YY}^{1/2}g\right\|^2+\left\|\Sigma_{YY}^{1/2}g\right\|^2-\left\|W_{ZY}\Sigma_{YY}^{1/2}g\right\|^2$$

$$=\underline{\left\|\Sigma_{ZZ}^{1/2}f-W_{ZY}\Sigma_{YY}^{1/2}g\right\|^2}+\left\langle g,\left(\Sigma_{YY}-\Sigma_{YY}^{1/2}W_{YZ}W_{ZY}\Sigma_{YY}^{1/2}\right)g\right\rangle$$

This part can be arbitrary small
by choosing *f*.

Relation with conditional covariance

Theorem (FBJ'06, Sun et al. '07)

X, Y, Z : random variables on Ω_X, Ω_Y, Ω_Z (resp.).

(H_X, k_X), (H_Y, k_Y), (H_Z, k_Z) : RKHS defined on Ω_X, Ω_Y, Ω_Z (resp.).

Assume

$$H_Z + \mathbf{R} \text{ : dense in } L^2(P_Z)$$

then,

$$\langle g, \Sigma_{YX|Z} f \rangle = E\big[Cov[g(Y), f(X) \mid Z]\big] \qquad (\forall f \in H_X, g \in H_Y)$$

– *c.f.* for Gaussian variable

$$a^T V_{XY|Z} b = Cov[a^T X, b^T Y \mid Z]$$

(not dependent on the value of z)

– Sketch of the proof for the simpler case of $X = Y$ and $f = g$,

i.e. $\langle g, \Sigma_{YY|Z} g \rangle = E\left[Var[g(Y) \mid Z] \right]$

Lemma $Var[Y] = Var_X\left[E_{Y|X}[Y \mid X] \right] + E_X\left[Var_{Y|X}[Y \mid X] \right]$

$$\langle g, \Sigma_{YY|Z} g \rangle = \inf_{f \in H_Z} Var\left[g(Y) - f(Z) \right]$$

$$= \inf_{f \in H_Z} \left\{ Var\left[E\left[g(Y) - f(Z) \mid Z \right] \right] + E\left[Var\left[g(Y) - \underbrace{f(Z)}_{\text{const.}} \mid \underline{Z} \right] \right] \right\}$$

$$= \inf_{f \in H_Z} Var\left[\underbrace{E\left[g(Y) \mid Z \right]}_{\in L^2(P_Z)} - f(Z) \right] + E\left[Var\left[g(Y) \mid Z \right] \right]$$

$$= 0 + E\left[Var\left[g(Y) \mid Z \right] \right]$$ (by denseness assumption)

Conditional Independence

<u>Theorem (FBJ04, Sun et al 07)</u>

Under (A),

$$\Sigma_{YX|Z} = O \quad \Leftrightarrow \quad P_{YX} = E_Z\left[P_{Y|Z} \otimes P_{X|Z}\right]$$

where $E_Z\left[P_{Y|Z} \otimes P_{X|Z}\right]$ is a probability on $\Omega_X \times \Omega_Y$ defined by

$$E_Z\left[P_{Y|Z} \otimes P_{X|Z}\right](B \times A) = \int P_{Y|Z}(B \mid Z = z) P_{X|Z}(A \mid Z = z) dP_Z(z)$$

With p.d.f.

$$E_Z\left[P_{Y|Z} \otimes P_{X|Z}\right](A \times B) = \int \int_A p_{X|Z}(x \mid z) d\mu_1(x) \int_B p_{Y|Z}(y \mid z) d\mu_2(y) dP_Z(z)$$

Remark: The assertion $P_{YX} = E_Z\left[P_{Y|Z} \otimes P_{X|Z}\right]$ is weaker than the conditional independence $P_{YX|Z} = P_{Y|Z} \otimes P_{X|Z}$

c.f. for Gaussian variables

$$V_{YX|Z} = O \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z$$

– Proof of $\Sigma_{YX|Z} = O \quad \Rightarrow \quad P_{YX} = E_Z\left[P_{Y|Z} \otimes P_{X|Z}\right]$

$\Sigma_{YX|Z} = O$ means $E\left[Cov[g(Y), f(X) \mid Z]\right] = 0$

$$\Rightarrow \quad E\left[E[g(Y)f(X) \mid Z]\right] = E\left[E[g(Y) \mid Z]E[f(X) \mid Z]\right]$$

$$\Rightarrow \quad E_{P_{XY}}[g(Y)f(X)] = E_{E_Z[P_{Y|Z} \otimes P_{X|Z}]}[g(Y)f(X)] \qquad \forall f \in H_X, g \in H_Y$$

Under (A), by approximating the index function $I_{A \times B}(x, y)$

$$P_{YX} = E_Z\left[P_{Y|Z} \otimes P_{X|Z}\right]$$

Characterization of conditional independence

<u>Theorem</u>

Define the augmented variable $\tilde{X} = (X, Z)$ and define a kernel on $\Omega_X \times \Omega_Z$ by

$$k_{\tilde{X}} = k_X k_Z$$

Under (A),

$$\Sigma_{Y\tilde{X}|Z} = O \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z$$

$$\Sigma_{Y\tilde{X}|Z} = O \quad \Leftrightarrow \quad \Sigma_{\tilde{Y}X|Z} = O \quad \Leftrightarrow \quad \Sigma_{\tilde{Y}\tilde{X}|Z} = O \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z$$

proof)

$$\Sigma_{Y[X,Z]|Z} = O \quad \Rightarrow \quad p(x, y, z') = \int p(x, z' \mid z) p(y \mid z) p(z) dz$$

where $p(x, z' \mid z) = p(x \mid z) \delta(z' - z)$

$$\Rightarrow \qquad p(x, y, z') = p(x \mid z') p(y \mid z') p(z')$$

i.e. $\quad p(x, y \mid z') = p(x \mid z') p(y \mid z')$

Normalized Cond. Covariance

- Normalized conditional cross-covariance operator

Definition

$$W_{YX|Z} = \Sigma_{YY}^{-1/2}\Sigma_{YX|Z}\Sigma_{XX}^{-1/2} = \Sigma_{YY}^{-1/2}\left(\Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}\right)\Sigma_{XX}^{-1/2}$$

More rigorously,

$$W_{YX|Z} \equiv W_{YX} - W_{YZ}W_{ZX}$$

Recall: $\Sigma_{YX} = \Sigma_{YY}^{1/2}W_{YX}\Sigma_{XX}^{1/2}$

– Conditional independence

Under the assumption (A),

$$W_{Y\tilde{X}|Z} = O \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z$$

Conditional Dependence Measure

– HS Normalized Conditional Independence Criteria

$$HSNCIC = \left\| W_{\tilde{X}\tilde{Y}|Z} \right\|_{HS}^{2}$$

$$HSNCIC = 0 \qquad \Leftrightarrow \qquad X \perp\!\!\!\perp Y \mid Z$$

– Hilbert-Schmidt norm of an operator

$A : H_1 \to H_2$ operator on a Hilbert space

A is called Hilbert-Schmidt if for complete orthonormal systems $\{\varphi_i\}$ of H_1 and $\{\psi_j\}$ of H_2

$$\sum_j \sum_i \langle \psi_j, A\varphi_i \rangle^2 < \infty.$$

Hilbert-Schmidt norm is defined by

$$\|A\|_{HS}^2 = \sum_j \sum_i \langle \psi_j, A\varphi_i \rangle^2$$

c.f. Frobenius norm of a matrix

Kernel-free Expression

<u>Theorem</u>

Assume

P_{XY} and $E_Z\left[P_{Y|Z} \otimes P_{X|Z}\right]$ have density $p_{XY}(x,y)$ and $p_{X\perp Y|Z}(x,y)$, resp.

$H_Z + \mathbf{R}$ and $H_X \otimes H_Y + \mathbf{R}$ are dense in $L^2(P_Z)$ and $L^2(P_X \otimes P_Y)$, resp.

W_{YX} and $W_{YZ} W_{ZX}$ are Hilbert-Schmidt.

Then,

$$\| W_{YX|Z} \|_{HS}^2 = \iint \left(\frac{p_{XY}(x,y) - p_{X\perp Y|Z}(x,y)}{p_X(x) p_Y(y)} \right)^2 p_X(x) p_Y(y) dx dy$$

In the special case of $Z = \phi$

$$\| W_{YX} \|_{HS}^2 = \iint \left(\frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} - 1 \right)^2 p_X(x) p_Y(y) dx dy$$

- Kernel-free expression, though the definitions are given by kernels!

– Kernel-free value is reasonable as a “measure” of dependence. *c.f.* If unnormalized operators are used, the measures do depend on the choice of kernel (HSIC, Gretton et al. ALT2005)

– In the unconditional case,

$$\text{HS-NIC} = \| W_{YX} \|_{HS}^2$$

is equal to the mean square contingency, which is one of the popular measures of dependence.

– In the conditional case, if we use the augmented variables

$$\| W_{\tilde{Y}\tilde{X}|Z} \|_{HS}^2$$
$$= \iint \left(\frac{p_{XYZ}(x,y,z) - p_{X|Z}(x|z)\, p_{Y|Z}(y|z)\, p_Z(z)}{p_{XZ}(x,z)\, p_{YZ}(y,z)} \right)^2 p_{XZ}(x,z)\, p_{YZ}(y,z)\, dx dy dz$$

(conditional mean square contingency)

– Key idea of the proof

By the eigendecomposition of Σ_{XX} and Σ_{YY} , we have CONS $\{\varphi_i\}$ of H_X and $\{\psi_j\}$ of H_Y such that

$$\Sigma_{XX}\varphi_i = \lambda_i \varphi_i, \quad \Sigma_{YY}\psi_j = \nu_j \psi_j \qquad (\lambda_i \ge 0, \nu_j \ge 0)$$

Define

$$\tilde{\varphi}_i = \frac{\varphi_i - E[\varphi_i]}{\sqrt{\lambda_i}}, \quad \tilde{\psi}_j = \frac{\psi_j - E[\psi_j]}{\sqrt{\nu_i}}$$

By the denseness assumption, $\{1\} \bigcup \{\tilde{\varphi}_i \tilde{\psi}_j\}_{i,j}$ is CONS of $L^2(P_X \otimes P_Y)$

$$\sum_{i,j} \left\langle \psi_j, W_{YX}\varphi_i \right\rangle^2 = \sum_{i,j} \left\langle \Sigma_{YY}^{-1/2}\psi_j, \Sigma_{YX}\Sigma_{XX}^{-1/2}\varphi_i \right\rangle^2 = \sum_{i,j} \left\langle \frac{\psi_j}{\sqrt{\nu_j}}, \Sigma_{YX}\frac{\varphi_i}{\sqrt{\lambda_i}} \right\rangle^2$$

$$= \sum_{i,j} E_{XY}\left[\tilde{\psi}_j(Y)\tilde{\varphi}_i(X)\right]^2 = \sum_{i,j} \left(\tilde{\psi}_j(Y)\tilde{\varphi}_i(X), \frac{p_{XY}}{p_X p_Y} \right)^2_{L^2(P_X \otimes P_Y)}$$

$$= \left\| \frac{p_{XY}}{p_X p_Y} \right\|^2_{L^2(P_X \otimes P_Y)} - 1$$

etc.

Empirical Measures

- Empirical estimation is straightforward with the kernel method.
- Inversion → regularization: $\Sigma_{XX}^{-1} \quad \to \quad (\Sigma_{XX} + \varepsilon I)^{-1}$
- Replace the covariances in $W_{YX} = \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}$ by the empirical ones given by the data $\Phi_X(X_1),\ldots,\ \Phi_X(X_N)$ and $\Phi_Y(Y_1),\ldots,\ \Phi_Y(Y_N)$

$$HSNIC_{emp} = \mathrm{Tr}\left[R_X R_Y\right]$$ (dependence measure)

$$HSNCIC_{emp} = \mathrm{Tr}\left[R_{\ddot{X}} R_{\ddot{Y}} - 2R_{\ddot{X}} R_{\ddot{Y}} R_Z + R_{\ddot{X}} R_Z R_{\ddot{Y}} R_Z\right]$$

(conditional dependence measure)

where $R_{\ddot{X}} \equiv G_{\ddot{X}}\left(G_{\ddot{X}} + N\varepsilon_N I_N\right)^{-1}$ etc.

- $\text{HSNIC}_{\text{emp}}$ and $\text{HSNCIC}_{\text{emp}}$ give kernel estimates for the mean square contingency and conditional mean square contingency, resp.

Relation with Other Measures

- Mutual Information

$$MI(X,Y) = \iint p_{XY}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\,d\mu_X(x)d\mu_Y(y)$$

- MI and HSNIC

$$\boxed{HSNIC(X,Y) \le MI(X,Y)}$$

$\because$)

$$\begin{aligned} HSNIC &= \iint p_{XY}(x,y)\left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} - 1\right)d\mu_1(x)d\mu_2(y) \\ &\le \iint p_{XY}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\,d\mu_1(x)d\mu_2(y) \quad = MI \end{aligned}$$

$$(\log z \le z - 1)$$

– Mutual Information:
  - Information-theoretic meaning.
  - Estimation is not straightforward for continuous variables. Explicit estimation of p.d.f. is difficult for high-dimensional data.
    – Parzen-window is sensitive to the band-width.
    – Partitioning may cause a large number of bins.
  - Some advanced methods: e.g. k-NN approach (Kraskov et al. 2004, Ku&Fine 2005).

– Kernel method:
  - Explicit estimation of p.d.f. is not required;

    the dimension of data does not appear explicitly, but it is influential in practice.
  - Kernel / kernel parameters must be chosen.

Statistical Consistency

Theorem (FGSS2007)

Assume that $W_{YX|Z}$ is Hilbert-Schmidt, and the regularization coefficient satisfies

$$\varepsilon_N \to 0 \qquad N^{1/3}\varepsilon_N \to \infty.$$

Then,

$$\left\| \hat{W}_{YX|Z}^{(N)} - W_{YX|Z} \right\|_{HS} \to 0 \qquad (N \to \infty)$$

In particular,

$$\left\| \hat{W}_{YX|Z}^{(N)} \right\|_{HS} \to \left\| W_{YX|Z} \right\|_{HS} \qquad (N \to \infty)$$

i.e. $\mathrm{HSNCIC}_{\mathrm{emp}}$ ($\mathrm{HSNIC}_{\mathrm{emp}}$) converges to the population value HSNCIC (HSNIC, resp).

Choice of Kernel

- How to choose a kernel?
  - Empirical estimates still depend on the choice of kernels.
  - For unsupervised problems, such as independence measures, there are no theoretically reasonable methods.

  - Some heuristic methods which work:
    - Heuristics for Gaussian kernels

$$\sigma = \mathrm{median}\left\{ \left\| X_i - X_j \right\| \ | \ i \neq j \right\}$$

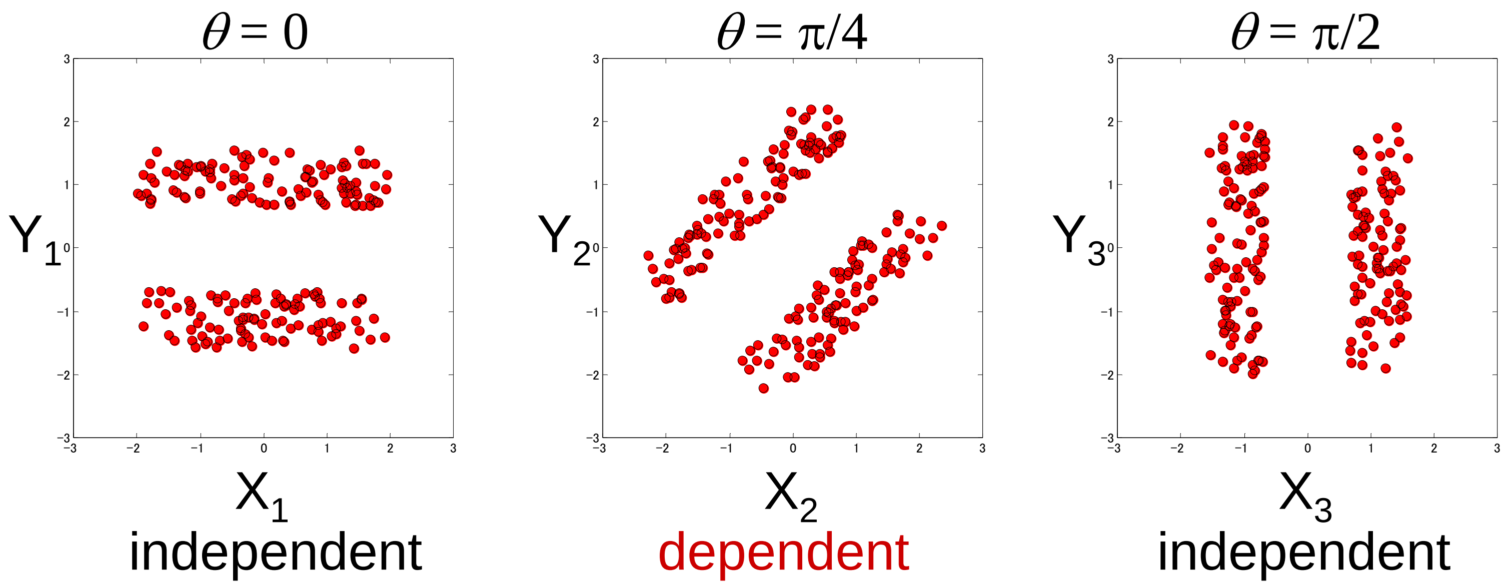
    - Speed of asymptotic convergence

$$\lim_{N \to \infty} Var\left[N \times HSNIC_{emp}^{(N)} \right] = 2 \left\| \Sigma_{XX} \right\|_{HS}^2 \left\| \Sigma_{YY} \right\|_{HS}^2 \quad \text{under independence}$$

Compare the bootstrapped variance and the theoretical one, and choose the parameter to give the minimum discrepancy.

Application to Independence Test

- Toy example



They are all uncorrelated, but dependent for $0 < \theta < \pi/2$

N = 200.
Permutation test is used.

indep. → more dependent

Angle	0.0	4.5	9.0	13.5	18.0	22.5
HSIC (Median)	93	92	63	5	0	0
HSIC (Asymp. Var.)	93	44	1	0	0	0
HSNIC ($\varepsilon = 10^4$, Median)	94	**23**	**0**	**0**	**0**	**0**
HSNIC ($\varepsilon = 10^6$, Median)	92	**20**	**1**	**0**	**0**	**0**
HSNIC ($\varepsilon = 10^8$, Median)	93	**15**	**0**	**0**	**0**	**0**
HSNIC (Asymp. Var.)	94	**11**	**0**	**0**	**0**	**0**
MI (#NN = 1)	93	62	11	0	0	0
MI (#NN = 3)	96	43	0	0	0	0
MI (#NN = 5)	97	49	0	0	0	0

acceptance of independence out of 100 tests (significance level = 5%)

Cond. Independence Test

- Permutation test with the kernel measure

$$T_N = \left\| \hat{\Sigma}_{YX|Z}^{(N)} \right\|_{HS}^2 \qquad \text{or} \qquad T_N = \left\| \hat{W}_{YX|Z}^{(N)} \right\|_{HS}^2$$

– If Z takes values in a finite set $\{1, \ldots, L\}$,

set $A_\ell = \{i \mid Z_i = \ell\} \ (\ell = 1, \ldots, L)$,

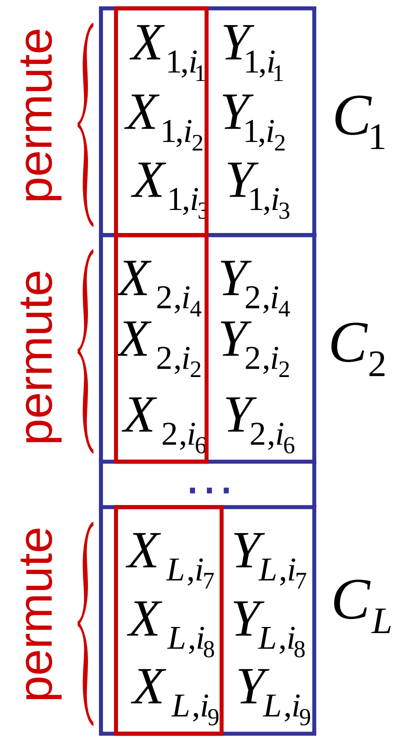
otherwise, partition the values of Z into L subsets $C_1, \ldots, C_L$, and set

$$A_\ell = \{i \mid Z_i \in C_\ell\} \quad (\ell = 1, \ldots, L).$$

– Repeat the following process B times: ($b = 1, \ldots, B$)

1. Generate pseudo cond. independent data $D^{(b)}$ by permuting X data within each A_ℓ.
2. Compute $T_N^{(b)}$ for the data $D^{(b)}$.

$\longrightarrow$ Approximate null distribution under cond. indep. assumption



– Set the threshold by the $(1-\alpha)$-percentile of the empirical distributions of $T_N^{(b)}$.

Kernel Method for Causality of Time Series

Causality by conditional independence

– Nonlinear extension of Granger causality

X is NOT a cause of Y if

$$p(Y_t \mid Y_{t-1},\ldots,Y_{t-p}, X_{t-1},\ldots,X_{t-p}) = p(Y_t \mid Y_{t-1},\ldots,Y_{t-p})$$

$\Leftrightarrow$

$$Y_t \perp\!\!\!\perp X_{t-1},\ldots,X_{t-p} \mid Y_{t-1},\ldots,Y_{t-p}$$

– Kernel measures for causality

$$HSNCIC = \left\| \hat{W}^{(N-p+1)}_{\tilde{Y}\mathbf{X_p}|\mathbf{Y_p}} \right\|^2_{HS}$$

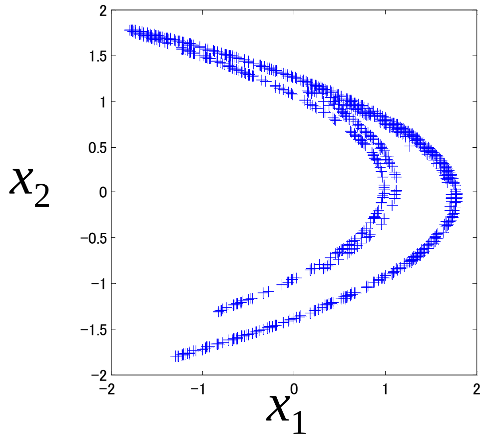
$$\mathbf{X}_{\mathrm{p}} = \{(X_{t-1,} X_{t-2}, \cdots, X_{t-p}) \in \mathbf{R}^p \mid t = p+1,\ldots,N\}$$
$$\mathbf{Y}_{\mathrm{p}} = \{(Y_{t-1,} Y_{t-2}, \cdots, Y_{t-p}) \in \mathbf{R}^p \mid t = p+1,\ldots,N\}$$

Example: Causality of Time-Series

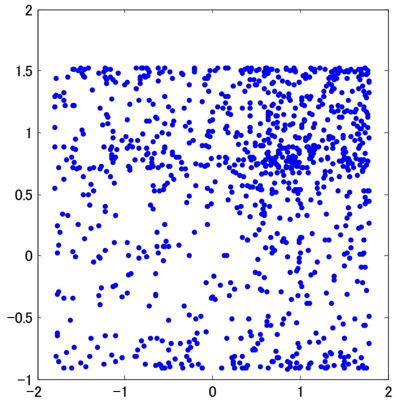
- Coupled Hénon map
  - X, Y:

$$\begin{cases} x_1(t+1) = 1.4 - x_1(t)^2 + 0.3x_2(t) \\ x_2(t+1) = x_1(t) \end{cases}$$

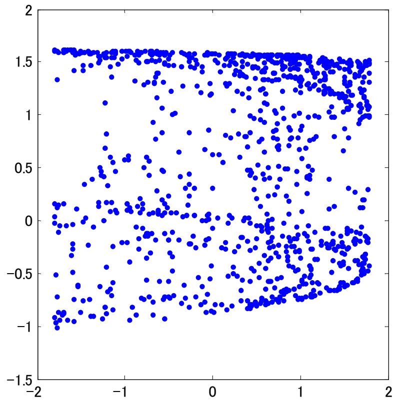
$$\begin{cases} y_1(t+1) = 1.4 - \left\{\underline{\gamma x_1(t)} y_1(t) + (1-\gamma) y_1(t)^2\right\} + 0.1y_2(t) \\ y_2(t+1) = y_1(t) \end{cases}$$



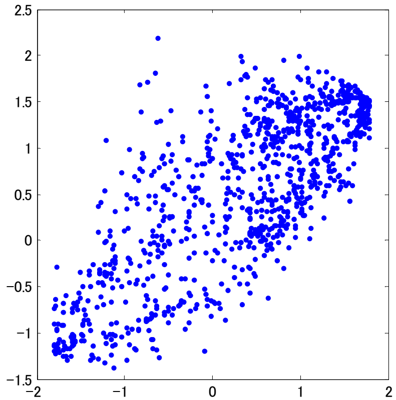
x_1-y_1



$\gamma = 0$



$\gamma = 0.25$



$\gamma = 0.8$

■ Causality of coupled Hénon map

- X is a cause of Y if $\gamma > 0$. $$Y_t \not\perp\!\!\!\perp X_{t-1},\dots,X_{t-p} \mid Y_{t-1},\dots,Y_{t-p}$$
- Y is not a cause of X for all γ. $$X_t \perp\!\!\!\perp Y_{t-1},\dots,Y_{t-p} \mid X_{t-1},\dots,X_{t-p}$$
- Permutation tests for non-causality with $HSNCIC = \left\| \hat{W}^{(N-p+1)}_{\ddot{Y}\mathbf{X_p}|\mathbf{Y_p}} \right\|^2_{HS}$

N = 100

$x_1 - y_1$	H$_0$: Y_t is not a cause of X_{t+1}							H$_0$: X_t is not a cause of Y_{t+1}						
γ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.0	0.1	0.2	0.3	0.4	0.5	0.6
HSNCIC	94	88	81	63	86	77	62	97	0	0	0	0	0	0
Granger	92	96	95	90	90	94	93	96	92	85	45	13	2	3

1-dimensional independent noise is added to $X(t)$ and $Y(t)$.

HSNCIC	97	96	93	85	81	68	75	96	0	0	0	0	0	0

Number of times accepting H$_0$ among 100 datasets (α = 5%)

Concluding Remarks

- Kernel dependence measures
  - The normalized (conditional) covariance on RKHS gives kernel-free measures of dependence in population.
  - The Gram matrix expression gives the p.d.-kernel estimate of the (conditional) mean square contingency.
  - Comparably reliable methods for conditional independence test.
- Future directions
  - More empirical studies
  - More theory on kernel choice
  - Application to causal inference (Sun et al., 2007).